Investigation, Visualization, and Interpretation of Large Scientific Data Sets

Dr. Brian Circelli, Paul Adams
U.S. Army Engineer Research and Development Center
Major Shared Resource Center
Vicksburg, MS

Dr. Joseph Werne, Dr. Michael Gourlay,
Dr. Christian Meyer, and Dr. Chris Bizon
Colorado Research Associates Division
NorthWest Research Associates, Inc.
Boulder, CO

Overview

- Identify the difficulties associated with large data sets
- Proposed alternatives to manipulating large data sets
- Scientific visualization of large data sets
- ERDC MSRC visualizations
- Conclusions

Identify the Difficulties Associated with Large Data Sets

- Storage and transfer of data on the order of hundreds of gigabytes to several terabytes.
  - Currently the data is transferred over a high-speed network to a mass storage computer and migrated to robotic tape storage.
- Extraction of useful information contained within the large data sets.
  - 3D scientific visualization is used for qualitative understanding & interpretation.

Data Transfer Limitations

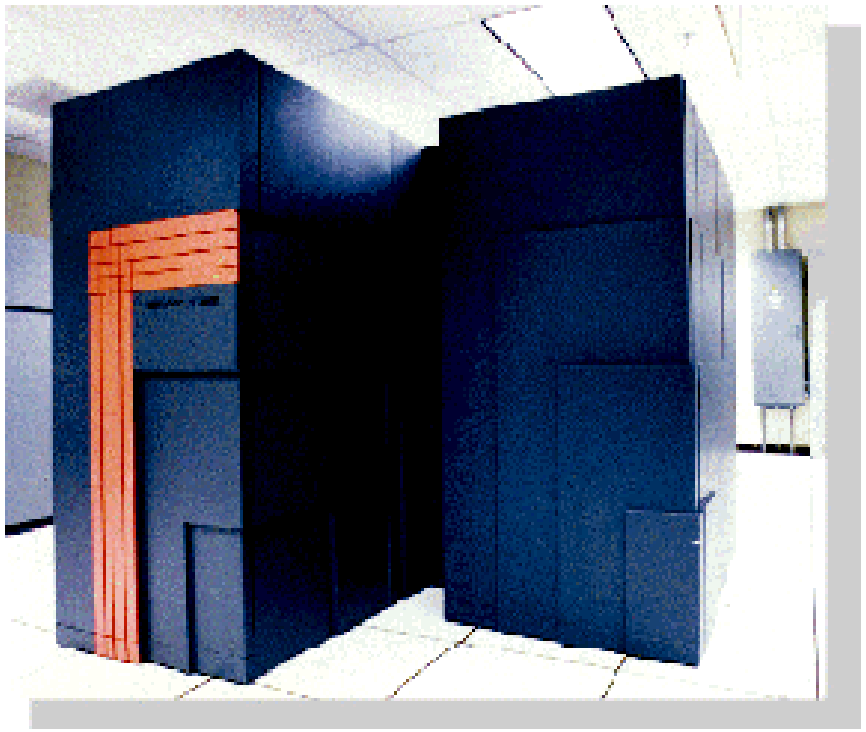
- Theoretical Network Transfer Time of 6 Terabytes
  - HiPPI
    - 800 Megabits/sec transfer rate
    - 17.5 hours
    - Internal network connection
  - ATM OC-12
    - 622 Megabits/sec transfer rate
    - 22.5 hours
    - Internal and DREN network connection
  - Fast Ethernet
    - 100 Megabits/sec transfer rate
    - 5.8 days
  - Ethernet
    - 10 Megabits/sec transfer rate
    - 58 days
- Contention with other network traffic lowers the theoretical transfer rate and raises the time to transfer the data

Storage Limitations

- Limited disk space on HPC machines
  - Competition for disk storage space necessitates purging cycles to be imposed.
  - Disk space may be unavailable when needed.
- Cost of adding more disks
  - Adding disk storage space is expensive and encounters long lead times.

ERDC Cray T3E

- 544 Total PE
- 600+ Gflops peak performance
- 520 Application PE
  - 600 MHz PE
  - 256 Mbytes Memory/PE
- 700 Gbytes of /tmp disk space
- HiPPI Network Interface
  - Connects to HAFS
  - Connects to MSF



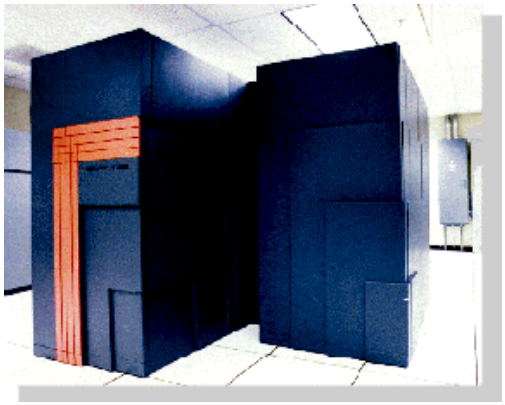
ERDC Cray MSF

- Cray J916
  - 64 MegaWords
  - 206 Gbytes disk space
- Attaches to three StorageTek (STK) 9310 Robotic Tape Silos
  - 500+ Terabytes storage
- HiPPI Network Interface

ERDC SGI HAFS

- High Availability File Server (HAFS)
- SGI Origin 2000
  - 32 CPU
  - 195 MHz CPU
- 800 Gbytes disk space
  - Fibre Channel
  - RAID 5

Standard Migration Path of Data Files



Cray T3E - Jim

HIPPI

SGI Origin 2000 - HAFS

Cray J90 - MSF

To Remote Backup

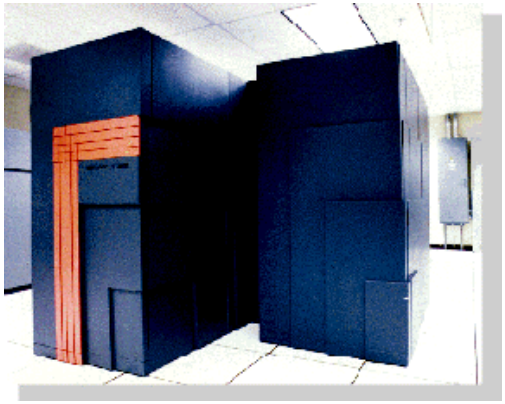
Benefits of the ERDC Standard Migration Path

- Remote Backup copy of data is archived
- Two copies of data are written on MSF for archival purposes
- Disk space on HAFS is larger than on MSF

Storage Problems Encountered with the ERDC Standard Migration Path

- MSF
  - 206 Gbytes of disk cache were too small and filled up quickly.
- HAFS
  - Spillover from MSF filled up HAFS disks
  - HAFS continually sent data back to MSF to be migrated
- T3E
  - Spillover from HAFS filled up T3E disks
- All HPC systems affected by spillover

Modified Migration Path



Cray T3E - Jim

HiPPI

Cray J90 - MSF

Benefits of the ERDC Modified Migration Path

- Spillover limited to MSF/T3E
- No contention for HAFS resources
- No contention for remote backup resources
- Allowed for direct control of migrating data

Problems Encountered with the ERDC Modified Migration Path

- Modified Migration Path has no remote backup
- Disk space on MSF is still limited

Proposed alternatives to manipulating large data sets

- Perform data post-processing analysis & visualization on the supercomputer using a Client/Server visualization package.
  - Data remains local to the supercomputer, which provides terabytes of disk space.
  - Available disk space is only temporary & is subject to purging.
  - Disk space required for post-processing & visualization may be unavailable due to the sharing of computer resources with competing users.

Proposed Alternatives to Manipulating Large Data Sets (cont.)

- Perform data post-processing analysis, data byte scaling on the supercomputer & then transfer the results to a home-site Sci-Vis machine for visualization.
  - Provides a more tangible method of data set manipulation.
    - byte scaling provides a factor of 4 reduction in data set size
    - narrowing the visualization to a selected region of interest within the overall flow field (For the work presented here, a factor of 3 reduction in data set size was achieved.)
    - compression of data via gzip provided a factor of 5 to 6 reduction in data set size

Scientific Visualization of Large Data Sets

- Software tools that take advantage of advanced hardware architecture design.
  - 2D & 3D internal hardware texture mapping.
- Ogle: a 3D vector & scalar scientific visualization tool based on openGL
  - Well documented, supported and easy to use.
  - Possesses multifunctional capabilities, including:
    - rendering data set volumes
    - locating streamline paths
    - plotting vector field arrows
    - plotting isosurfaces
  - Capable of reading compressed data files

ERDC MSRC Visualizations

- Airborne Laser Challenge Project II
  - Data post-processing analysis, data byte scaling, & visualization of more than 6 terabytes of data.
  - Consumed more than 50,000 CPU hours combined on the NAVO MSRC & ERDC MSRC Cray -T3E supercomputers.
  - Collaborative visualizations were presented during the CFD Session C of the UGC.
  - Representative images of the vortex tube behavior during the break down of a KH vortex.

ERDC MSRC Visualizations (cont.)

- Airborne Laser Visualizations

ERDC MSRC Visualizations (cont.)

- Vortex Dynamics and Late-Wake Turbulence in Stratification and Shear Challenge Project.
  - Animations & scientific visualizations of late-wake turbulence.
  - Assisted in the understanding & interpretation of data generated from the large-scale DNS.
  - Collaborative visualizations were presented during the Challenge Projects Session C of the UGC.
  - Representative images of 3D coherent pancake vortices that exist in a zero momentum density stratified flow.

ERDC MSRC Visualizations (cont.)

- Wake Turbulence Visualizations

Conclusions

- Fundamental difficulties associated with large scientific data sets include:
  - data storage
  - data transport
  - data analysis
  - data visualization
  - data interpretation
- Solutions to the difficulties hinge upon finding ways to reduce data set size without degrading accuracy and include:
  - data byte scaling
  - narrowing interest to only a subset of the overall flow domain
  - data set compression via utilities such as gzip
- Ogle is a viable software tool for scientific visualization of large data sets whose capabilities include:
  - rendering 3d data set volumes
  - locating streamline paths
  - plotting vector field arrows